Results and Evaluation of Hungarian Nominal WordNet v1.0

Márton Miháltz and Gábor Prószéky

MorphoLogic
Késmárki u. 8, Budapest, 1118 Hungary
Email: mihaltz@morphologic.hu, proszeky@morphologic.hu

Abstract. This paper presents recent results of the ongoing project aimed at creating the nominal database of the Hungarian WordNet. We present 9 different automatic methods, developed for linking Hungarian nouns to WN 1.6 synsets. Nominal entries are obtained from two different machine-readable dictionaries, a bilingual English-Hungarian and an explanatory monolingual (Hungarian). The results are evaluated against a manually disambiguated test set. The final version of the nominal database is produced by combining the verified result sets and their intersections when confidence scores exceeded certain threshold values.

1 Introduction

The project started in 2000, with the aim of creating a Hungarian nominal WordNet ontology with semi-automatic methods [6]. Our basic strategy was to attach Hungarian entries of a bilingual English-Hungarian dictionary to the nominal synsets of Princeton WordNet, version 1.6 (WN, [4]), following the so-called expand approach [7]. This way, the synsets formed by the Hungarian nouns can inherit the WN semantic relations. In order to achieve this, we used heuristic methods, developed partly by previous similar projects [1,2], and partly by us, which rely on information extracted from several machine-readable dictionaries (MRDs). This approach relies on the assumption that nominal conceptual hierarchies, which describe the world, would be similar across English and Hungarian languages to a degree which is sufficient for producing a preliminary version of our WordNet.

2 Machine-Readable Dictionaries Used

We used two different MRDs to assist the heuristics which disambiguate the Hungarian nouns against Princeton WordNet synsets. The *MoBiDic* bilingual English-Hungarian electronic dictionary contains 17,700 Hungarian nominal entries, corresponding to 12,400 English equivalents covered in WordNet 1.6. These Hungarian nouns serve as the basis of the attachment procedure.

The other MRD we used is an electronic version of *the Magyar Értelmező Kéziszótár* (EKSz, [3]) monolingual explanatory dictionary. It covers over 42,000 nominal headwords, whose different senses correspond to over 64,000 different definitions. We used these definitions to gain semantic information in order to assist the heuristics that disambiguate Hungarian nouns against WN synsets via their English translations in the bilingual dictionary.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 175–180.

3 Methods

The bilingual dictionary provides 1.71 English translations on average for each Hungarian nominal headword. These English translations correspond to 2.16 WordNet synsets on average. We implemented several heuristic methods in order to accomplish the automatic disambiguation of Hungarian nouns against the candidate WN synsets.

3.1 Methods Relying on the Bilingual Dictionary

The first group of heuristics was developed by Atserias et al for the Spanish WordNet project [1]. These heuristics rely on information found in the connections between Hungarian and English words in the bilingual dictionary, and between English headwords and corresponding synsets in WN.

- MONOSEMIC METHOD: if an English headword is monosemous with respect to WN (belongs to only one synset), then the corresponding Hungarian headword is linked to the synset.
- VARIANT METHOD: if a WN synset contains two or more English words that each has only one translation to the same Hungarian word, it is linked to this synset.
- INTERSECTION METHOD: links a Hungarian headword to all synsets sharing at least two of its English translations.

A fourth kind of heuristic depends on morpho-semantic information found in the Hungarian side of the bilingual dictionary. A number of Hungarian headwords in the bilingual dictionary are endocentric (noun + noun) compounds, which have the property that the second segment of the compound defines the semantic domain of the whole word. For example, the compound *hangversenyzongora* (‘grand piano’) can be analysed as *hangverseny*+*zongora* (‘concert’+‘piano’), where the second segment, *zongora* serves as the DERIVATIONAL HYPERNYM noun of the compound. This piece of semantic information can be used with the modified conceptual distance formula (Section 3.2) in order to select a synset from the candidate ones.

3.2 Methods Relying on the Monolingual Explanatory Dictionary

The nominal definitions of the EKSz monolingual explanatory dictionary were POS-tagged and morphologically analyzed using the Humor analyzer [5].Using this information to recognize morpho-syntactic patterns, we were able to identify genuses, or hypernym words in 53,500 definitions, synonyms (10,500 definitions), plus holonyms (826 definitions) and meronyms (584 definitions).

Part of the acquired semantic information was used for the attachment of Hungarian nouns in the following way:

- SYNONYMS: the synset is chosen from the ones available for all the translations of the headword, which contains the greatest number of the synonyms’ English translations.
- HYPERNYMS: for those cases where both the headword and the corresponding acquired hypernym have English translations, the headword is disambiguated against WordNet using a modified version of the conceptual distance formula, developed by Atserias et al. [1], shown in Figure 1.

$$dist'(w_1, w_2) = \min_{\substack{c_{1i} \in w_1 \\ c_{2j} \in w_2 \\ depth(c_{1i}) < depth(c_{2j})}} |path(c_{1i}, c_{2j})|$$

Fig. 1. The simplified conceptual distance formula is applied to the pairs of English translations of a Hungarian noun and its hypernym. The formula returns two concepts (WN synsets) representing words which are closest to each other in the WN hypernym hierarchy

A third heuristic depends on the LATIN equivalents available for about 1,500 EKSz headwords, mostly covering various animal or plant species, taxonomic groups, diseases etc. Since WN also contains most of these Latin words in different synsets, these could be used to attach the EKSz headwords in a straightforward way.

Performance of all the individual methods relying on the bilingual and monolingual dictionaries is shown in Table 1.

Table 1. Performance of each method: number of Hungarian nouns and WN synsets covered, and number Hungarian noun-WN synset connections

Method	Hungarian nouns	WN 1.6 synsets	Connections
Mono	8 387	5 369	9 917
Intersection	2 258	2 335	3 590
Variant	164	180	180
DerivHyp + CD	1 869	1 857	2 119
EKSz synonyms	927	707	995
EKSz hypernyms + CD	5 432	6 294	9 724
EKSz Latin equivalents	1 697	838	848

3.3 Methods for Increasing Coverage

In those cases where the identified hypernyms or synonyms had no English translations, we used two methods to gain a related hypernym word that has a translation and hence can be used to disambiguate with the aid of the modified conceptual distance formula.

The first method was to look for derivational hypernyms of the synonyms or hypernyms, using the methods described above. Since hypernymy is transitive, the hypernym of the headword's hypernym (or synonym) will also be a hypernym.

The other method looks up the hypernym (or synonym) word as an EKSz an entry, and if it corresponds to only one definition (eliminating the need for sense disambiguation), then the hypernym word identified there is used, if it is available (and has English equivalents). These two methods provided a 9.2% increase in the coverage of the monolingual methods.

Table 2 summarizes the results of all the automatic methods used on different sources in the automatic attachment procedure.

Table 2. Total figures for the different types of methods

Type of Methods	Hungarian nouns	WN 1.6 synsets	Connections
Bilingual only	10 003	7 611	13 554
Monolingual	7 643	7 380	10 901
Monoling. + incr. cov. 1–2	8 343	8 199	12 185
Total	**13 948**	**12 085**	**22 169**

4 Validation and Combination of Results

In order to validate the performance of the automatic methods, we constructed a manual evaluation set consisting of 400 randomly selected Hungarian nouns from the bilingual dictionary, corresponding to 2 201 possible WN synsets through their English translations. Two annotators manually disambiguated these 400 words, which meant answering 2 201 yes-no questions asking whether a Hungarian word should be linked to a WN synset or not. Inter-annotator agreement was 84.73%. In the cases where the two annotators disagreed, a third annotator made the final verdict.

We first validated the different individual methods against the evaluation set. The results are shown in Table 3.

Table 3. Precision and recall on the evaluation set, plus coverage of all Hungarian entries in the bilingual dictionary for the individual attachment methods, in descending order of precision. The Latin method is not included, because for the most part it covers terminology not covered by the general vocabulary of the evaluation set

Method	Precision	Recall	Coverage
Variant	92.01%	50.00%	0.50%
Synonym	80.00%	39.44%	8.00%
DerivHyp	70.31%	69.09%	17.50%
Incr. cov. 1.	67.65%	46.94%	7.50%
Mono	65.15%	55.49%	69.25%
Intersection	58.56%	35.33%	17.50%
Incr. cov. 2.	58.06%	28.57%	6.00%
Hypernym	48.55%	41.71%	49.25%

Atserias et al [1] and Farreres et al [2] describe a method of manually checking the intersections of results obtained from different sources. They determined a threshold (85%) that served as an indication of which results to include in their preliminary WN. Then drawing upon the intuition that information discarded in the previous step might be valuable if it was confirmed by several sources, they checked the intersections of all pairs of the discarded result sets. This way, they were able to further increase the coverage of their WNs without decreasing the previously established confidence score of the entire set.

We used a similar approach. We decided to use two thresholds, 70% and 65%, creating the bases for two versions of the final nominal WN (*min65* and *min70*). The first set included results from the VARIANT, SYNONYM and DERIVHYP methods, the second contained these

plus the results from the INC. COV. 1 method and MONO methods. Both sets also included the results from the LATIN methods, as manual inspections estimated its precision to be fairly high (over 80%). Table 5 shows the figures for the base sets.

The next step was to validate the intersections of all the pairs of results not included in the previous step. The scores for the best-performing combinations are presented in Table 4.

Table 4. Precision, recall and coverage of intersections of sets not included in the base sets

Intersections of methods	Precision	Recall	Coverage
Inc. cov. 2. & Hypernym	95.78%	50.00%	1.50%
DerivHyp & Inc. cov. 2.	94.64%	80.03%	1.00%
DerivHyp & Intersection	92.20%	75.10%	0.75%
Inc. cov. 2. & Intersection	88.14%	90.00%	0.50%
Inc. cov. 2. & Mono	87.50%	70.00%	2.00%
DerivHyp & Mono	84.38%	87.10%	8.00%
Hypernym & Mono	71.91%	52.46%	21.00%
DerivHyp & Hypernym	70.97%	66.67%	7.25%
Hypernym & Intersection	67.86%	30.16%	6.25%

For the two final versions of the Hungarian nominal WN 1.0, we combined the min70 and min65 base sets with intersection sets having precision score over 70% and 67%, respectively (Tables 5 and 4).

Table 5. Overall results for the two versions of Hungarian nominal WordNet v1.0, with their constituting base and intersection sets

Result set	#Words	#Synsets	#Connections	Precision
min70 base	2 445	2 170	2 722	76.14%
min70 additional intersections	7 183	6 142	8 579	76.70%
min70 final set	**7 927**	**6 551**	**9 635**	**75.38%**
min65 base	12 275	11 597	20 439	65.11%
min65 additional intersections	3 110	2 698	3 431	66.91%
min65 final set	**12 839**	**12 004**	**22 169**	**63.35%**

5 Conclusions, Further Work

We used several automatic methods to attach Hungarian nominal headwords of a bilingual dictionary to WN 1.6 synsets. The various heuristics were validated against a manually disambiguated set, and from their combinations we produced two versions of the nominal database, having estimated precisions of 63 and 75 percent, with different numbers of words covered.

There are two ways to further enrich our initial nominal WN. On the one hand, to increase its coverage, we will apply the methods which proved to be most successful (VARIANT, SYNONYM, DERIVHYP) on new sources—additional bilingual dictionary modules, dictionaries with multi-word phrases, thesauri etc.

On the other hand, in order to increase the confidence of the existing result sets, a completely manual checking of the links between WN 1.6 synsets and Hungarian nouns will be necessary. This will have to rely on strict guidelines, which will be based on the pilot work disambiguating the entries in the evaluation set.

We have also applied for funding to support work on the further extension of our core Hungarian WN. This would include: revising the entire WN from a point of view independent of the English Princeton WN, adding databases for remaining other parts of speech, and connecting our WN to the EuroWordNet [8] framework.

References

1. Atserias, J., S., Climent, X., Farreres, G., Rigau, H., Rodríguez: Combining multiple methods for the automatic construction of multilingual WordNets. Proc. of Int. Conf. on Recent Advances in Natural Language Processing, Tzigov Chark (1997).
2. Farreres, X., G., Rigau, H., Rodriguez: Using WordNet for building Wordnets. Proc. of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal (1998).
3. Juhász, J., I., Szöke, G. O. Nagy, M. Kovalovszky (eds.): Magyar Értelmező Kéziszótár. Akadémiai Kiadó, Budapest: (1972).
4. Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller: Introduction to WordNet: an on-line lexical database. Int. J. of Lexicography 3 (1990) 235–244.
5. Prószéky, Gábor: Humor: a Morphological System for Corpus Analysis. Language Resources and Language Technology, Tihany (1996) 149–158.
6. Prószéky, G. M. Miháltz: Automatism and User Interaction: Building a Hungarian WordNet. Proc. of the Third International Conference on Language Resources and Evaluation, Las Pa mas de Gran Canaria, Spain (2002).
7. Vossen, P.: Right or Wrong. Combining lexical resources in the EuroWordNet project. Proceedings of Euralex-96, Goetheborg (1996).
8. Vossen, P. (eds): EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht (1998).